

## BIG DATA: OPPORTUNITIES & CHALLENGES

SONAL CHAWLA, NEETIKA GOSWAMI & PRIYANKA NARWAL

Research Scholar, Department of CSE, PDM College of Engineering, Bahadurgarh, Haryana, India

### ABSTRACT

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. This review paper discusses types of big data, its opportunities & challenges and some applications of big data.

**KEYWORDS:** Big Data, Storage, 3v's, It Applications

### INTRODUCTION

The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte- or larger-sized datasets with high-velocity and different structures. Big data<sup>[1]</sup> has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. P. Zikopoulos, J.J. Berman defined big data as characterized by three Vs: volume, variety, and velocity. The terms volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges. IDC also defined big data technologies as "new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis."

3 Vs of Big Data:

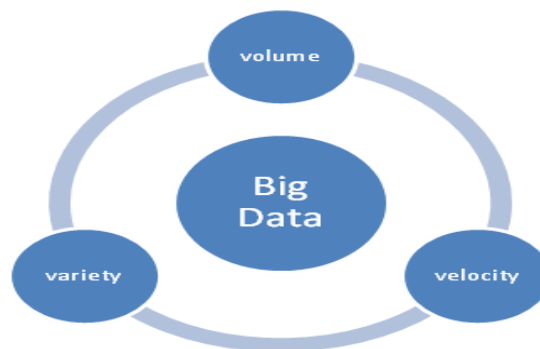


Figure 1: 3Vs of Big Data

- **Volume** refers to the amount of all types of data generated from different sources and continue to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis. Laurila et al. provided a unique collection of longitudinal data from smart mobile devices and made

this collection available to the research community. The aforesaid initiative is called mobile data challenge motivated by Nokia. Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data challenge produced an interesting result similar to that in the examination of the predictability of human behavior patterns or means to share data based on human mobility and visualization techniques for complex data<sup>[2]</sup>.

- **Variety** refers to the different types of data collected via sensors, smartphones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data.
- **Velocity** refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources.

## TYPES OF BIG DATA

Big data is divided in three types describing the purpose for which they are used.

- **Structured Data** is the type that would fit neatly into a standard Relational Data Base Management System, RDBMS, and lend itself to that type of processing. Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data[3].
- **Semi-structured Data** is that which has some level of commonality but does not fit the structured data type. For example, Web logs, Social media and E-commerce.
- **Unstructured Data** is the type that varies in its content and can change from entry to entry. For example, Pictures, Video Editing, Productivity (office documents) and Geological data. These data cannot easily be separated into categories or analyzed numerically.

## OPPORTUNITIES & CHALLENGES

Big data technologies are maturing to a point in which more organizations are prepared to pilot and adopt big data as a core component of the information management and analytics infrastructure. Big data, as a compendium of emerging disruptive tools and technologies, is positioned as the next great step in enabling integrated analytics<sup>[5]</sup> in many common business scenarios.

As big data wends its inextricable way into the enterprise, information technology (IT) practitioners and business sponsors alike will bump up against a number of challenges that must be addressed before any big data program can be successful<sup>[8]</sup>. Five of those challenges are:

- Uncertainty of the Data Management Landscape

There are many competing technologies, and within each technical area there are numerous rivals. Our first

challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.

- The Big Data Talent Gap

The excitement around big data applications seems to imply that there is a broad community of experts available to help in implementation. However, this is not yet the case, and the talent gap poses our second challenge.

- Getting Data into the Big Data Platform

The scale and variety of data to be absorbed into a big data environment can overwhelm the unprepared data practitioner, making data accessibility and integration our third challenge.

- Synchronization Across the Data Sources

As more data sets from diverse sources are incorporated into an analytical platform, the potential for time lags to impact data currency and consistency becomes our fourth challenge.

- Getting Useful Information out of the Big Data Platform

Lastly, using big data for different purposes ranging from storage augmentation to enabling high-performance analytics is impeded if the information cannot be adequately provisioned back within the other components of the enterprise information architecture, making big data syndication our fifth challenge.

## APPLICATIONS

Big data has increased the demand of information management specialists<sup>[11]</sup>. It is estimated that one third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications.

### Government

The use and adoption of Big Data within governmental processes is beneficial and allows efficiencies in terms of cost, productivity, and innovation. That said, this process does not come without its flaws. Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome.

### International Development

Research on the effective usage of information and communication technologies for development (also known as ICT4D) suggests that big data technology can make important contributions but also present unique challenges to International development. Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management.

### Manufacturing

Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for manufacturing. Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability.

## Private Sector

### Retail

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.

### Retail Banking

- FICO Card Detection System protects accounts world-wide.
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.

### Real Estate

- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

### Technology

- eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse
- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- Facebook handles 50 billion photos from its user base.

## CONCLUSIONS

Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. With Big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

## REFERENCES

1. S.Kaisler, F.Armour, J.A.Espinosa, W.Money, BigData: Issues and Challenges Moving Forward, System Sciences (HICSS), 2013, in: Proceedings of the 46th Hawaii International Conference on, IEEE, 2013, pp.995–1004.
2. N.Leavitt, Storage challenge: where will all that big data go? Computer 46(2013)22–25.
3. K.Strauss, D.Burger, What the future holds for solid-state memory, Computer 47(2014)24–31.
4. L.Chih-Wei, H.Chih-Ming, C.Chih-Hung, Y.Chao-Tung, An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation, Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, 2013, pp.463–468.
5. L.Chang, R.Ranjan, Z.Xuyun, Y.Chi, D.Georgakopoulos, C.Jinjun, Public Auditing for Big Data Storage in Cloud Computing – a Survey, Computational Science and Engineering (CSE), 2013 IEEE 16th International

- Conferenceon, 2013, pp.1128–1135.
6. M. Cox, D. Ellsworth, Managing Big Data For Scientific Visualization, ACM Siggraph, 1997.
  7. O. Kwon, N. Lee, B.Shin, Data quality Q7 management, data usage experience and acquisition intention of big data analytics, *Int.J.Inf. Manag.* (2014).
  8. K. Singh, S. C. Guntuku, A. Thakur, C.Hota, Big data analytics framework for peer-to-peer botnet detection using random forests, *Inf. Sci.*(2014).
  9. J.L.Schnase, D.Q.Duffy, G.S.Tamkin, D.Nadeau, J.H.Thompson, C.M.Grieg, M.A.McInerney, W.P.Webster, MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud- enabled Climate Analytics-as-a-Service, *Computers, Environment and Urban Systems*, (2014).
  10. B.K.Tannahill, M.Jamshidi, System of systems and bigdata analytics – bridging the gap, *Comput.Electr.Eng.*4 (2014)2–15.
  11. J.Abawajy, Symbioses of Big Data and Cloud Computing: Opportunities & Challenges, (2013).
  12. S. Aluru, Y.Simmhan, A special issue of journal of parallel and distributed computing: scalable systems for big data management and analytics, *J.Parallel Distrib.Comput.*73 (2013)896.
  13. S. Hipgrave, Smarter fraud investigations with big data analytics, *Netw.Secur.*2013 (2013)7–9.

